CS512 F24 Project Proposal

Saurabh Rajput – A20523129, Arpita Jadhav – A20523353

Main Paper

Title: "Segment Anything"

Authors: Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick
Publication Details: Meta AI Research, 2023. Available on arXiv.
Link: Segment Anything

We have chosen Option 2: Modify a significant component of the paper to improve the original work.

Title: Enhancing the Segment Anything Model (SAM) with Improved Prompt Encoding and Multi-Scale Feature Fusion for Domain-Specific Segmentation

Objective:

The Segment Anything Model (SAM), developed by Meta AI, was designed with the objective of creating a generalizable segmentation model that can handle a variety of segmentation tasks with minimal fine-tuning. The model is built on the principle of providing promptable segmentation, where users can input prompts (e.g., points, boxes, text) to guide the segmentation process across different images and contexts. SAM's flexible prompt-based interface enables its use in multiple applications, from medical imaging to autonomous driving.

While SAM has achieved significant milestones, particularly in general-purpose segmentation, it faces challenges when applied to domain-specific tasks like medical imaging and object segmentation in autonomous vehicles. SAM's prompt encoder, for example, does not fully leverage user inputs for detailed segmentation in these specialized domains. Additionally, the model struggles with handling scale variance that is the ability to correctly segment objects of different sizes and resolutions and lacks the fine-tuning necessary for adapting to specific domains with unique data characteristics.

Therefore, the objective of this project is to build upon the foundational work of the original SAM by addressing its key limitations. We plan to enhance the performance of the Segment Anything Model (SAM) by improving its prompt encoder and incorporating multi-scale feature fusion. This aims to increase accuracy and efficiency in domain-specific segmentation tasks such as medical imaging and autonomous driving.

Challenges:

Limited Prompt Interpretation: The current prompt encoder may not fully leverage user inputs for precise segmentation, particularly in complex or specialized contexts.

Scale Variance Handling: SAM sometimes struggle with objects of varying sizes due to the lack of multi-scale feature processing, affecting detailed segmentation tasks.

Domain Adaptability: Without fine-tuning, SAM might not achieve optimal performance in specialized domains where data characteristics differ from general datasets.

Approach:

We plan to implement two significant modifications to SAM and lastly domain-specific fine-tuning:

A. Enhanced Prompt Encoder

Integration of Attention Mechanisms:

How: We plan on incorporating multi-headed self-attention layers within the prompt encoder to allow the model to focus on relevant regions based on user inputs. Attention mechanisms can improve the model's ability to interpret prompts, leading to more accurate and context-aware segmentation.

Learned Prompt Embeddings:

How: developing trainable embeddings for different prompt types (points, boxes, masks) to better integrate them with image features. Learned embeddings capture nuanced relationships between prompts and image content, enhancing segmentation precision.

Hierarchical Input Encoding:

How: encoding user inputs at multiple scales, allowing the model to capture both fine-grained details and global context. Hierarchical encoding enables the model to understand objects at various levels, improving segmentation of complex structures.

B. Multi-Scale Feature Fusion

Incorporation of Feature Pyramid Network (FPN):

How: integrating an FPN into SAM's architecture to extract and fuse features at multiple resolutions as multi-scale features allow the model to handle objects of different sizes, enhancing segmentation accuracy across scales.

Cross-Resolution Attention:

How: implementing attention mechanisms that operate across different feature scales to combine detailed local information with global context. This facilitates better information flow between scales, improving the model's ability to segment objects with complex spatial structures.

C. Domain-Specific Fine-Tuning

Fine-Tuning on Specialized Datasets:

How: Adapt the modified SAM by training it on domain-specific datasets for example, medical images, autonomous driving scenes. Fine-tuning will allow the model to learn domain-specific features and improve performance in specialized tasks.

Implementation Plan:

Modify the Prompt Encoder:

- Integrate attention layers and learned embeddings.
- Test the enhanced encoder with various prompts.

Incorporate Multi-Scale Feature Fusion:

- Embed an FPN into the architecture.
- Implement cross-resolution attention mechanisms.

Fine-Tune the Model:

- Train the modified SAM on selected datasets.
- Adjust training parameters to optimize performance.

Evaluation:

- We plan to compare the performance of the enhanced SAM with the original model using metrics like Mean Intersection over Union (mIoU) and inference time.
- Analyze improvements in segmentation accuracy and efficiency.

Data

Datasets we are exploring:

Medical Imaging:

Medical Segmentation Decathlon (MSD):

Description: A collection covering various anatomies with diverse imaging modalities.

LIDC-IDRI:
Description: Lung CT scans for nodule detection and segmentation.

Autonomous Driving:

Cityscapes Dataset:
Description: Urban street scenes with pixel-level annotations.

KITTI Dataset:
Description: Real-world driving data for object detection and segmentation.

Team Member Responsibilities

Saurabh Rajput:

- Work on the enhancement of the prompt encoder by integrating attention mechanisms and developing learned embeddings.
- Document all code modifications with detailed inline comments.
- Fine-tune the enhanced model on the selected datasets.
- Collaborate on testing and debugging the modified components.

Arpita Jadhav:

- Implement multi-scale feature fusion into the SAM architecture, including cross-resolution attention mechanisms.
- Prepare and preprocess the domain-specific datasets for training and evaluation.
- Analyze performance improvements and generate visualizations for the report.

References

Kirillov, A., et al. "Segment Anything." arXiv preprint arXiv:2304.02643 (2023).

Lin, T.-Y., et al. "Feature Pyramid Networks for Object Detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017): 2117-2125.

Vaswani, A., et al. "Attention is All You Need." Advances in Neural Information Processing Systems 30 (2017): 5998-6008.

Simpson, A. L., et al. "A Large Annotated Medical Image Dataset for the Development and Evaluation of Segmentation Algorithms." arXiv preprint arXiv:1902.09063 (2019).

Cordts, M., et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016): 3213-3223.

He, K., et al. "Mask R-CNN." Proceedings of the IEEE International Conference on Computer Vision (2017): 2961-2969.

Howard, A., et al. "Searching for MobileNetV3." Proceedings of the IEEE/CVF International Conference on Computer Vision (2019): 1314-1324.